Steven S. Gouveia
Doutorado em Neurofilosofia da Mente, investigador do Instituto de Filosofia da Universidade do Porto

"PODEMOS ESTAR PERANTE A ÚLTIMA CRIAÇÃO DO SER HUMANO"

É uma realidade inegável e está à vista de todos: a revolução digital entrou numa nova fase e a Inteligência Artificial (IA) veio para ficar. Com este poderoso avanço tecnológico, muitas são as questões e dilemas de natureza Ética que se colocam e que são determinantes para perceber o real alcance da IA no futuro e nas nossas vidas. Steven S. Gouveia, filósofo português, doutorado em Neurofilosofia da Mente, tem-se dedicado a refletir e a escrever sobre o tema, sendo o título do seu livro mais recente *"Thinking the New World: Conversations on Artificial Intelligence"*. Atualmente investigador do Instituto de Filosofia da Universidade do Porto, Steven S. Gouveia lidera um projeto de seis anos sobre a Ética da Inteligência Artificial em medicina, em parceria com as Universidades de Yale, Helsínquia e Exeter. "Estamos perante uma fase da Humanidade que requer o máximo de esforço e 'expertise' ética para considerar todas as potenciais implicações em desenvolver este tipo de tecnologias altamente complexas que podem revolucionar todos os aspetos do que significa Ser Humano", afirma. A regulação é, pois, fundamental.

Vivemos tempos de fulgurante progresso tecnológico e digital, que está a desvendar novas avenidas do conhecimento, ainda que não saibamos bem até onde nos podem levar. Como estudioso destas matérias, que convocam a uma reflexão mais filosófica, como analisa o momento que vivemos?

Penso que vivemos em tempos filosoficamente extraordinários. Sendo doutorado em Filosofia, há sempre um debate sobre a inutilidade desta profissão quando confessamos aos nossos entrequeridos a decisão de dedicar uma vida à área mais antiga do conhecimento humano. Ouvimos os conselhos sábios dos nossos pais e amigos para decidir estudar outro assunto, um assunto mais 'prático', mais simples, que possa realmente ter utilidade e garantir um salário estável. Ora, contrariamente a esse conselho amável de preocupação, de facto, estamos a começar a entrar a todo o vapor na *"Idade da Inteligência Artificial"*, como indica o título do livro que editei em 2019, que conta com a participação de alguns dos maiores especialistas em inteligência artificial do mundo – como Ben Goertzel, o criador do Robot Sofia, Daniel Dennett, um dos filósofos mais influentes do mundo, entre muitos outros. Por isso mesmo, a área da Filosofia que lida com os problemas levantados por essa nova era está cheia de trabalho de investigação a ser realizado, dada a proliferação imensa de várias tecnologias em vários setores da sociedade. Dito isto, é preciso considerar que estamos perante uma tecnologia que poderá ser diferente de todas as outras na história da Humanidade: podemos estar perante a última criação do ser humano, dado que, pelo menos em teoria, estamos perante a possibilidade de criar uma inteligência artificial que poderá multiplicar a sua inteligência exponencialmente com total autonomia. Tal cenário poderá ser um grande risco existencial, dado que essa super-inteligência poderá olhar para nós, seres humanos, como nós olhamos para as formigas: como meros seres, mais ou menos desprezíveis, sem relevância moral alguma. Por isso mesmo, estamos perante uma fase da humanidade que requer o máximo de esforço e expertise ética para considerar todas as potenciais implicações em desenvolver este tipo de tecnologias altamente complexas que podem revolucionar todos os aspetos do que significa Ser Humano.

As gerações contemporâneas cresceram com livros, filmes ou outros formatos artísticos que, na área da ficção científica, imaginaram cenários que se assemelham, em maior ou menor medida, aos tempos que vivemos, refletindo sobre conceitos que hoje são realidades, como por exemplo a Inteligência Artificial (IA) ou a Robótica. O futuro traz consigo imensas incógnitas. Devemos estar otimistas ou mais apreensivos? As máquinas vão controlar os humanos?

A resposta honesta é dizer que tudo dependerá de nós, humanos. É verdade que a cultura pop tem influenciado de diversas formas o olhar filosófico da população sobre diversos temas, desde os apocalipses zombies à IA. O importante é aprendermos com os erros do passado – o que olhando para a história da Humanidade, é algo difícil de acontecer – e não repetirmos os processos relacionados com os casos da genética humana ou a energia bélica nuclear. O que aconteceu, nestes dois casos, foi ter havido o desenvolvimento da tecnologia numa primeira instância e, só depois, se ter começado a pensar eticamente as variadas implicações da tecnologia. Ora, tal revela um viés falacioso contra o qual é preciso lutar com todas as nossas forças: o viés de colocar o desenvolvimento tecnológico em primeiro lugar e, só numa segunda instância, considerar a ética dessa tecnologia. Tal passa principalmente pelo papel dos governos darem

"[Existe] um viés falacioso contra o qual é preciso lutar com todas as nossas forças: o viés de colocar o desenvolvimento tecnológico em primeiro lugar e, só numa segunda instância, considerar a ética dessa tecnologia. Tal passa principalmente pelo papel dos governos darem a relevância necessária à Ética e à expertise ética, naturalmente fornecida por eticistas e filósofos morais especializados."

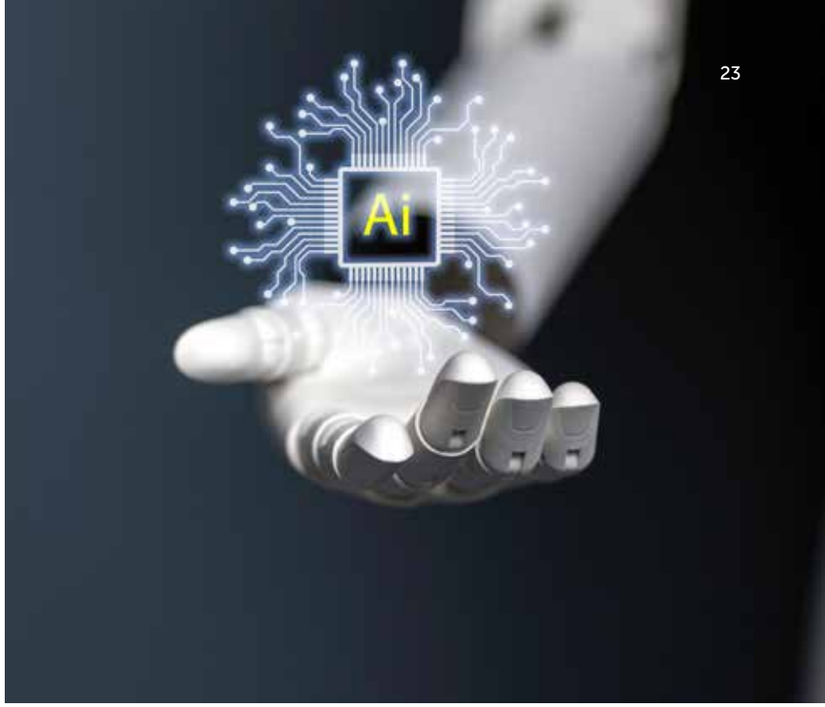
a relevância necessária à Ética e à expertise ética, naturalmente fornecida por eticistas e filósofos morais especializados. Diferente desta expertise é a situação com que lidamos atualmente, em que a Ética é reduzida a meros pseudo-regulamentos internos, em que as comissões de ética de várias instâncias, como hospitais, tribunais, universidades, etc., são constituídas por várias profissões, como médicos, juristas, sociólogos, políticos, entre muitas outras, menos a profissão e *expertise* que está literalmente na sua denominação: Ética. É muito raro, se não impossível, encontrar eticistas – pessoas formadas nessa área do saber filosófico – nas comissões de ética, o que já diz muito sobre o quão pouco preparada está a sociedade para lidar com os problemas ligados à IA. Penso ser absolutamente determinante corrigir esta situação se queremos evitar a tal repetição falaciosa da história.

Conceitos como a IA já são realidades que vieram para ficar. Como fica e que papel assume o chamado elemento humano tendo em conta todas estas evoluções tecnológicas e digitais?

O conceito de IA é, atualmente, um conceito que não refere nada de concreto no mundo, dado a proliferação tão excessiva de várias tecnologias que são referidas como IA, absolutamente dissemelhantes entre si. A melhor resposta a uma pergunta desta natureza é especificar em particular uma tecnologia e analisar as suas várias implicações. Por exemplo, desde maio de 2023 lidero um projeto de investigação de seis anos no Instituto de Filosofia da Universidade do Porto (com parceria com as Universidades de Yale, Helsínquia e Exeter) sobre a Ética da Inteligência Artificial na Medicina. O objetivo do projeto é compreender de que forma podemos desenvolver várias tecnologias médicas, sem que tal possa colocar em causa os princípios basilares da Medicina, como o Princípio da Confiança.

Como assim?

A influência da Inteligência Artificial na Medicina está a evoluir rapidamente na sociedade atual, baseada no pressuposto básico por trás dessa(s) tecnologia(s) produzirem práticas de saúde mais confiáveis, precisas, económicas e eficientes do que a chamada 'Medicina Tradicional', baseada no 'mero' raciocínio humano. O caminho alternativo é, então, a criação de algoritmos de diversas naturezas que irão auxiliar parcial ou totalmente os processos de tomada de decisão em contextos médicos. Contudo, é muito importante notar que a maioria dessas tecnologias são baseadas em diferentes tipos de dados e conteúdos complexos e multifacetados e que, por causa disso, a maioria desses sistemas de IA são puras "caixas negras" (black boxes): o especialista de saúde será capaz de entender as informações de entrada (*inputs*) do sistema e, depois, as informações de saída (*outputs*); porém, nunca terá acesso ao que acontece dentro do sistema. Assim, poderemos compreender os *inputs* e *outputs* de um determinado sistema, mas não conseguiremos explicar, do ponto de vista humano, por que um *input* específico está relacionado com um output específico e vice-versa, criando, deste modo, um processo opaco e obscuro em termos de explicação e compreensão. Ora, tal cenário levanta um problema ético enorme: por um lado, os médicos e especialistas de saúde terão a obrigação moral de confiar num sistema médico que é mais eficaz que eles próprios, embora não tenham qualquer compreensão do processo de decisão; e por outro, os pacientes terão de confiar num especialista médico que não nos pode oferecer uma explicação sobre o diagnóstico ou tratamento específico indicado pelo sistema de IA, que é uma "caixa negra". Para resolver este problema será preciso analisar as várias respostas possíveis que podem ser oferecidas: i) banir, simplesmente, qualquer sistema médico que tenha como base uma IA; ii) aceitar os benefícios da Medicina IA ignorando os seus malefícios; iii) desenvolver uma Inteligência Artificial Explicativa de segunda ordem que possa ser aplicada à IA e que possa trazer a compreensão que está em falta. Todas estas soluções têm problemas e vantagens e apenas uma reflexão ética profunda poderá compreender ao detalhe quais são os padrões éticos necessários para o desenvolvimento e uso destas tecnologias aplicadas à saúde, algo que espero desenvolver nos próximos anos. Convido desde já todos os leitores a aceder às várias atividades do projeto nowebsite https://trustaimedicine.weebly.com/



"Já temos várias evidências que a IA pode, muitas vezes, desenvolver vieses racistas ou sexistas em diversas tecnologias."

A IA pode ser tendenciosa? Legitimando, reproduzindo ou ampliando desigualdades ou preconceitos que já existem na sociedade? O que mostram outras revoluções tecnológicas históricas?

Já temos várias evidências que a IA pode, muitas vezes, desenvolver vieses racistas ou sexistas em diversas tecnologias. Por exemplo, um estudo que implementou uma tecnologia de machine learning para detetar cancro de pele usou 95% de dados de pacientes de pele branca para criar o seu modelo, o que levou a um viés estrutural contra pacientes de pele escura (cf. Zou, J. & Schiebinger. L. (2018) *"AI can be sexist and racist - it's time to make it fair"*, Nature, 559 (7714): 324-326.). O grande problema aqui é que a grande vantagem desta tecnologia – o facto de considerar processar milhões de dados sobre determinado assunto em questões de minutos ou segundos – é também o seu grande calcanhar de Aquiles ético. Porquê? Porque, por um lado, o mais realista são os dados, o mais realista será o algoritmo; contudo, o outro lado da moeda é que se a realidade produz os vieses racistas e sexistas que são uma realidade na nossa sociedade humana, então esses mesmos dados terão vieses e inclinações racistas e sexistas que serão altamente imorais se não forem corrigidos. O problema é que, novamente, esta tecnologia tem uma estrutura *black-box*, o que significa que é impossível identificar empiricamente um determinado dado e corrigi-lo. Há, contudo, vários mecanismos que podemos usar para tentar proceder a este tipo de correções: por exemplo, criar princípios éticos relacionados com a aquisição de dados, em que se exige que as bases de dados na sociedade sejam o mais diversificadas possível, mesmo que tal possa levar a um sistema um pouco menos eficaz do que poderia sê-lo em teoria.

Como responsabilizar algoritmos de IA por decisões erradas ou injustas? A nível ético, quais são os grandes desafios nestas áreas? No futuro, haverá alguma forma de implementar código de conduta ético-deontológica específica para a IA, de forma a garantir que seja desenvolvida e usada de forma responsável?

No meu livro *Homo "Ignarus"*, com prefácio de Peter Singer (2020, Minerva Editora), abordo a questão da responsabilidade na Inteligência Artificial. Tradicionalmente, a relação entre o artificial e o humano é fácil de ser caracterizada em termos normativos. Veja-se, por exemplo, um trabalhador de uma fábrica automóvel que opera uma máquina particular. O trabalhador é responsabilizado por qualquer consequência da operacionalização da máquina. Caso o trabalhador seja negligente e cause algum acidente, a pessoa, e somente a pessoa, será responsabilizada pelas consequências dessa negligência. Seria absurdo alguém considerar que a própria máquina deva ser, de algum modo, responsabilizada pelo ato. No máximo, caso haja um defeito da máquina, a responsabilidade pelas consequências da aplicação dessa máquina passa do operador para o seu fabricante. Assim, a ideia é que a moralidade é baseada num princípio em que algo é responsável – este algo pode ser qualquer entidade ou agente – se, e só se, possuir alguma propriedade mental como a consciência ou livre-arbítrio, que só os seres humanos têm e que a tecnologia parece não ter. No entanto, muitas tecnologias atuais já desenvolveram níveis de autonomia total com implicações para a realidade humana.

Que exemplos?

Por exemplo, os carros autónomos, que podem autonomamente conduzir-nos de um ponto A ao ponto de B. A máquina já não opera sob a égide de um agente humano, mas é capaz de decidir partes ou a totalidade de um curso de ação sem qualquer intervenção humana. Versões ainda mais avançadas podem começar por uma lista de regras fixada por programador, mas não ficar por aí: o algoritmo pode aprender com as suas operações e mudar as próprias regras (o que chamamos de machine learning). Ora, o ponto que devemos considerar é que a tese convencional de atribuição de responsabilidade não parece ser compatível com os avanços da tecnologia e o nosso sentido de justiça, e é exatamente isto a que chamamos de Lacuna de Responsabilidade. O que acabo por defender no meu livro é a ideia de recusar as abordagens éticas minimalistas baseadas no agente moral individual e consciente, dado que já não fazem sentido para pensar uma tecnologia que resulta de inúmeras interações entre muitos atores, que incluem os designers, programadores, utilizadores, software e hardware, redes, empresas, Estados, entre outros. Pelo contrário, devemos considerar uma abordagem de Responsabilidade Distribuída, que pode acomodar esta multiplicidade de agentes: é por isso que o conceito de Moralidade Distribuída é útil para a discussão do impacto das novas tecnologias da Inteligência Artificial, e é neste pressuposto que desenvolvo a minha solução ética para lidar com o problema da responsabilidade no meu livro "Homo Ignarus".

Muito recentemente, foi publicada uma carta aberta na qual se alerta para os riscos do uso acelerado da Inteligência Artificial (IA) e se admite que a corrida pelo desenvolvimento destes softwares está fora de controlo. Elon Musk, um dos principais instigadores do desenvolvimento acelerado da IA, é um dos subscritores. Eticamente, que responsabilidades podem/devem ser imputadas a estes 'empreendedores visionários' que estão a contribuir para uma tecnologia cujo impacto final é completamente desconhecido?

Acho que foi um exemplo interessante do quão a cultura pop influencia esta área: somos altamente influenciados por alguma cultura pop, como os filmes de Hollywood, em relação a esse assunto. Não estamos perto de ter uma tecnologia que pensa de forma independente. O caso do ChatGPT é claro: parece ser bom a pensar, mas quando começamos a fazer perguntas inteligentes, vemos que tal é apenas uma ilusão. Outro grande problema ligado a esta questão é que ainda não temos a menor ideia do que significa "Inteligência Humana": não há consenso científico ou filosófico sobre como definir esse conceito. Se não tivermos isso, é muito difícil usar a inteligência humana como medida para a inteligência da IA. Na minha opinião, para ter um sistema inteligente, será necessário desenvolver um sistema corpóreo, com um corpo completo, e garantir que esse corpo possa interagir com o ambiente e outras pessoas como os seres humanos interagem. Sem isso, penso que será muito difícil produzir inteligência real.

E como analisa as preocupações éticas relacionadas com o uso da IA em contexto militar e de defesa e segurança nacional? Como se pode garantir que este tipo de tecnologias usada de forma responsável nestas áreas tão sensíveis?

Não sou especialista neste tema em específico da Ética da IA. O que sabemos é que os cenários de guerra estão a mudar drasticamente com mais e mais utilização de tecnologia de guerra com capacidade de decisão autónoma. A seu favor, é argumentado que estas armas são mais eficazes e cometem menos erros que um operador humano dado que não tem medos, receios, nem vieses do tipo que os humanos possuem. Contudo, temos já dados suficientes para avaliar a suposta eficácia destes drones: um estudo recente indicava que embora oito inimigos fossem o alvo direto de um drone, mais de 1.600 pessoas acabaram por falecer por danos colaterais. Ora, a suposta eficácia parece-me estar em causa neste cenário (dados apresentados no meu livro de 2019 *"Reflexões Filosóficas: Arte, Mente e Justiça"*). Sobre este assunto, destaco o documentário internacional em que tive o prazer de ser host, onde entrevistei vários especialistas mundiais sobre diversos temas ligados a esta área, intitulado *"The Age of Artificial Intelligence: the Documentary"* (disponível gratuitamente no YouTube: https://www.youtube.com/watch?v=zMrt6ALaio0).

O Homem e a Máquina serão sempre dois polos inconciliáveis?

Novamente, essa resposta vai depender da tecnologia específica que estamos a abordar. No caso da utilização da IA na Medicina, acho que a melhor opção é o acordo de coexistência, à semelhança do que acontece com o Xadrez actual. Sabemos que o melhor jogador de xadrez humano contra o melhor *software* de xadrez artificial perderá facilmente. Mas a combinação de xadrez humano com o artificial pode vencer o xadrez artificial numa partida. Em outras áreas, acho que uma visão de substituição é provavelmente a mais interessante para libertar a humanidade de trabalhos chatos e aborrecidos que impedem o ser humano de ser ele mesmo, de ser um poeta à solta, como dizia Agostinho da Silva. Existem três formas específicas de considerar esta relação entre humanos e tecnologia: a primeira é considerá-la *in the loop*, onde o agente humano está a decidir ativamente sobre a decisão da tecnologia; uma segunda forma é ter uma estrutura *on the loop*, onde a IA pode tomar a decisão sozinha, mas tem um ser humano que supervisiona essa decisão e que tem sempre a última palavra; por fim, existe uma opção *out of the loop*, onde os humanos não têm nada a dizer sobre a decisão da IA. Acho que a maioria das decisões sobre a coexistência da tecnologia terão de cair nas duas primeiras categorias, e não na terceira.

Novas tecnologias implicam enormes recolhas de grandes quantidades de dados pessoais para tomar decisões, algo que levanta uma série de questões e preocupações relacionadas com a privacidade dos cidadãos. Como analisa esta questão?

Os usos de dados pessoais pelas novas tecnologias levantam questões e preocupações importantes em relação à privacidade, sendo que é necessário cumprir com os regulamentos atuais ou a desenvolver no futuro, como o Regulamento Geral de Proteção de Dados (RGPD) da União Europeia. É necessário também haver um maior controlo sobre a transparência do consentimento informado que, muitas vezes, não é tido com a seriedade que este deveria requerer – o caso das redes sociais é um exemplo claro disso, onde a maior parte das pessoas nem imagina que os seus dados pessoais estão a ser usados e vendidos a empresas. Deve haver também uma clara es-

pecificação de determinado uso dos dados pessoais: estes devem ser utilizados para o propósito indicado e não ter uma múltipla utilização. É também importante desenvolver tecnologia que possa garantir uma maior anonimização dos dados, através de técnicas de desidentificação de vários tipos que garantem a segurança dos dados contextuais e se focam apenas no dado absolutamente relevante para determinado serviço. Em suma, é preciso que a ideia de transparência e justiça esteja sempre presente em cada passo da estrutura tecnológica, seja na aquisição dos dados, na sua utilização ou processamento. Tal implica uma 'atitude' ética de todos os atores desta rede multifacetada, incluindo um papel ativo do Estado como regulador.

Qual a sua opinião com as preocupações relacionadas com esta evolução para um novo patamar da automação de empregos e outras questões sociais relacionadas ao uso da IA? Como é possível equilibrar o desenvolvimento de tecnologias que aumentam a eficiência com o impacto social potencial sobre as pessoas que podem ser afetadas negativamente pela automação/mecanização?

Penso que o problema mais grave da automação é a criação de níveis de desemprego históricos e incorrigíveis, no sentido de que a criação de novos empregos que possam surgir poderá não chegar para colmatar a criação do desemprego em massa. Ao contrário de outros períodos da história, estamos perante um desemprego que será massivo e que afetará milhões e milhões de pessoas. Porquê? Porque a tecnologia torna as despesas de uma empresa mais baixas. Torna a vida dos clientes mais simples e barata. Não há forma de contornar isso. Assim, é preciso pensar alternativas ao rendimento por trabalho e uma solução ou algo que considero fazer parte de uma solução maior que é avançar com a implementação do Rendimento Básico Incondicional (RBI) que pode ser financiado de variadas formas (por exemplo, taxando as grandes empresas tecnológicas que estão hoje livres de grande parte de contribuições). Esta ideia foi historicamente defendida por pensadores da esquerda à direita. O mais importante é que não podemos recusar esta ideia apenas porque sim, como temos visto a acontecer. É preciso investigar rigorosamente este tipo de políticas publicas e ver se realmente funcionam e não sermos influenciados pelas nossas intuições pré-concebidas. Por exemplo: uma objeção a esta ideia defende que as pessoas irão perder a motivação para ganhar dinheiro e irão gastar o seu dinheiro em droga e álcool. Ora, esta é uma objeção de carácter empírico: refere-se ao mundo real. Mas será que os primeiros estudos sobre a forma como as pessoas gastam o dinheiro de um RBI aponta para essa conclusão? A verdade é que não é de todo o que acontece: as pessoas aproveitam para voltar a estudar, para investir nas suas formações, para melhorarem as suas condições de vida. Claro, não estou a dizer que uma parte minoritária não possa efetivamente gastar a sua parte em drogas e álcool. Mas a exceção não faz a regra e não se pode colocar de lado o impacto positivo generalizado de uma ideia pela sua exceção negativa. Num dos cursos que leciono online e que estão disponíveis em formato gravação, abordo esta questão (disponível em https://stevensgouveia.weebly.com/curso-7.html).

> "É preciso que a ideia de transparência e justiça esteja sempre presente em cada passo da estrutura tecnológica, seja na aquisição dos dados, na sua utilização ou processamento. Tal implica uma 'atitude' ética de todos os atores desta rede multifacetada, incluindo um papel ativo do Estado como regulador."

Muitos setores – incluindo a saúde e os seguros – estão hoje dependentes da IA, inclusivamente no diagnóstico (de doenças e de risco, respetivamente). Que oportunidades se abrem à utilização deste tipo de tecnologia em áreas humanas tão sensíveis?

O uso da IA em áreas humanas sensíveis, como saúde e seguros, apresenta várias oportunidades para melhorar os respetivos resultados e transformar esses setores. Por exemplo, na saúde, como já indiquei, a IA tem o potencial de aumentar os recursos dos profissionais médicos, analisando grandes quantidades de dados do paciente, incluindo registos médicos, exames de imagem e informações genéticas. Já nos seguros, *chatbots* com IA podem lidar com consultas de rotina e tarefas administrativas, permitindo que os trabalhadores se concentrem em atividades mais complexas e criativas, sendo que essa automação pode levar, pelo menos em teoria, a uma maior eficiência, a custos mais reduzidos e a um ganho de produtividade. Podem também ser desenvolvidos algoritmos de AI para avaliar perfis de riscos através da variedade de dados de cada ser humano, o que pode ajudar a personalizar todo o sector dos seguros e saúde em geral, o que pode trazer benefícios em termos de custos mais baixos para os clientes. Claro, todos estes aspetos positivos levantam problemas éticos que já foram indicados por mim, mas também outros, como a questões ligadas à privacidade dos dados. A regulação deve garantir salvaguardas para assegurar a transparência e a proteção da privacidade e das informações confidenciais dos indivíduos, sem que tal possa prejudicar os mesmos nos serviços oferecidos.

"O problema mais grave da automação é a criação de níveis de desemprego históricos e incorrigíveis. Ao contrário de outros períodos da história, estamos perante um desemprego que será massivo e que afetará milhões e milhões de pessoas. Porquê? Porque a tecnologia torna as despesas de uma empresa mais baixas. Torna a vida dos clientes mais simples e barata. Não há forma de contornar isso."